基于自监督的场景去遮挡

Xiaohang Zhan[1], Xingang Pan[1], Bo Dai[1], Ziwei Liu[1], Dahua Lin[1], and Chen Change Loy[2]
[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong
[2]Nanyang Technological University
[1]{zx017, px117, bdai, zwliu, dhlin}@ie.cuhk.edu.hk
[2]ccloy@ntu.edu.sg

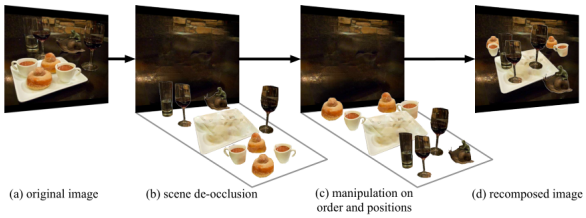

图 1: 场景去遮挡过程将一个图片分割，把其中杂乱且不完整的对象提取为单个完整对象的实体，可以操纵提取对象的顺序和位置以重新组合新的场景。

摘要—自然场景的理解是一个有挑战性的任务，尤其是当遇到多个物体互相遮挡的时候。这种理解障碍是由物体之间多变的排序和定位引起的。现有的场景理解方法只能对可视部分进行解析，导致对场景理解不完整和无条理的特性。在本文中，我们研究了场景去遮挡的问题，旨在恢复潜在的遮挡顺序，并对被遮挡物的不可视部分进行补全。我们首次尝试通过一个新的，统一的框架来恢复隐藏场景结构而非通过顺序和非模态标注。这一点是通过部分补全网络 (PCNet) 掩模部分 (M)，与内容部分 (C) 完成的，该网络可以学习如何以自监督方法恢复对象的掩模和内容部分。基于 PCNet-M 与 PCNet-C，我们设计了一种新的方法，通过逐步顺序恢复，非模态补全，内容补全三个阶段来完成场景去遮挡任务。对真实世界场景的试验证实了我们的方法相对其他方法有更好的表现。值得一提的是，我们使用自监督方法训练的模型在结果上可以媲美通过全监督方法。我们提出的场景去遮挡框架有助于许多其它的应用方向，包括高质量可控图像处理以及场景重建（如图1所示），以及现存模态掩模注释与非模态掩模注释之间的转换。项目主页：https://xiaohangzhan.github.io/projects/deocclusion/

I. 简介

场景理解是机器感知的基础内容之一。对真实世界下的场景，不论是什么内容，都包括不同顺序与位置下的多个物体，其中一部分物体被其他物体遮挡着。因此，场景理解系统应该能够用于模态感知 (解析可视区域)，非模态感知 [1]–[3](感知实体包括不可见部分的完整结构)。高级深度网络以及大规模标注数据集的出现促进了许多场景理解，目标检测 [4]–[7]，场景解析 [8]–[10]，实例分割任务 [11]–[14]。尽管如此，这些任务主要关注于模态感知，非模态感知至今仍很少有人探索过。

非模态感知的主要难点在于场景去遮挡，其主要包括恢复潜在遮挡顺序与补全被遮挡物不可视部分两个子任务。尽管人类视觉系统能够直觉地完成场景去遮挡，但对于机器来说，对遮挡部分的阐释很有挑战性。首先，对于遮挡其他物体的“遮挡物”与被遮挡的“被遮挡物”之间的关系，是非常复杂的。尤其是在有多个“遮挡物”与“被遮挡物”之间有复杂关系的情况下，单个“遮挡物”遮挡住了多个物体，且“被遮挡物”又被多个物体所遮挡，其间关系形成了一个复杂的遮挡图。其次，由于物体不同的种类，方向，位置，被遮挡物的边界是难以寻找的，没有简单的先验知识可以被用于寻找不可视部分的边界。

场景去遮挡问题一个可能的解决方案是用有遮挡顺序与非模态掩模（完整的示例掩模）的真实数据 (ground truth) 训练一个模型。这样的真实数据可以从人造数据 [15], [16] 或人工标注的真实世界数据 [17]–[19] 获得，但这两种方式都有其局限性。前者引入了伪造数据与测试中真实场景两者之间不可避免的域间隙，后者依赖于单个标注者的主观解释来标出被遮挡的界限，因此会产生误差，同时还需要不同标注者之间重复来减少噪声，因此费时费力。一个更加实用且可扩展方法是通过数据自身而非标注取学习场景去遮挡。

在这个工作中，我们提出了一个新的自监督框架取来处理真实世界中的场景去遮挡问题，而不使用人工标

注的遮挡顺序与非模态掩模。在缺乏真实场景的情况下，端对端的监督学习框架不再适用。因此，我们引入了一个特殊的概念：被遮挡物的部分补全。在部分补全的概念中，有两个核心原则，使得场景去障碍可以通过细监督方法完成。首先，补全一个被多个“遮挡物”遮挡的“被遮挡物”地过程可以被分解成一系列部分补全的序列，一次只对一个遮挡物进行处理。第二点，部分补全的学习可以通过进一步削减被遮挡物，同时训练一个网络去恢复未削减的被遮挡物来完成。我们证明了部分不全的方法对逐步完成被遮挡物是有效的且有助于对遮挡顺序的推理。

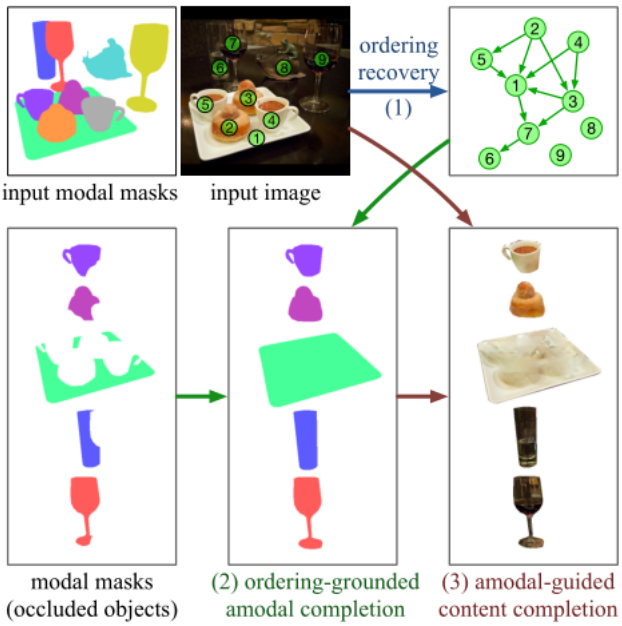

图 2: 已知输入图像与其模态掩模, 我们的框架以如下步骤逐步解决去遮挡问题：1) 预测不同物体之间顺序并构建有向图 2）基于有向图进行非模态补全 3）为被遮挡区域的非模态掩模填充内容。整个去遮挡过程通过两个网络：PCNet-C, PCNet-M 完成，这两个网络是在没有标注的顺序序列和非模态掩模上训练的。

部分补全过程是通过两个网络（掩模部分补全网络与内容部分补全网络）完成的，我们将他们分别缩写为 PCNet-M 与 PCNet-C。PCNet-M 是训练来部分地恢复被遮挡物对应于特定遮挡物的不可视部分的掩模，而 PCNet-C 是训练来用 RGB 内容对恢复的掩模的色彩填充。PCNet-M 和 PCNet-C 组成了解决去障碍框架的两个核心部分。

如图2所示，该框架以真实场景与其对应物体的掩模为基础，以已有模态分割的标注或预测作为输入。然后，我们的框架简化了三个步骤逐步执行。1）顺序恢复：给定一组相邻物体，其中一个可以遮挡住另一个物体，根据 PCNet-M 在保持遮挡物未更改的情况下部分完成被遮挡物掩模的原则，确定两个物体的角色。我们恢复所有相邻对，获得了一个能够表示所有物体顺序的一张有向图。2）非模态补全：对一个特定的被遮挡物，排序图表示了它所有的遮挡物。在此基础上，利用 PCNet-M，我们设计了一个非模态补全方法，将被遮挡物的模态掩模完善成为非模态掩模。3）内容补全：预测的非模态掩模显示了被遮挡物被遮挡的区域。我们向不可见区域通过使用 PCNet-C 提供 RGB 内容。通过这样一个渐进式的网络，我们把一个复杂的场景分割成了孤立的，完整的对象，产生了一个高度精确的有序图，同时可以将对象的顺序和位置进行后续处理以重新组合产生新场景，如图1所示。

我们做出的贡献如下：1）我们将场景去遮挡任务简化为三个子任务，分别为顺序恢复，非模态补全，内容补全。2）我们提出 PCNets 和一个新的推理方案来解决场景去障碍而非人工标注数据。但同时，我们观察了真实场景数据集上使用全监督方法与我们的方法的结果对比。3）这种自监督方法的特性显示了它可以为大型实例分割数据集 (如 KITTI [20],COCO [21]) 提供高精确度的顺序和非模态标注。4）我们的场景去障碍框架代表了一个新的支持真实场景操控充足技术，为图像编辑提供了一个新的维度。

II. 相关工作

顺序恢复：在无监督的方法中，吴等人 [22] 提出用通过用对象模板充足场景的方法恢复顺序。然而，他们只在小体积数据上证明了该系统。Tighe 等人 [23]，在训练集上建立了一个类间的先验遮挡矩阵并最小化二次规划在测试中排序，但先验遮挡矩阵忽略了真实场景的复杂性。还有其它的工作 [24], [25] 基于额外的深度信息，但在遮挡推理中深度信息并不可靠，举例来说，如果一张纸放在桌子上，他们的深度信息并无差别。这些工作所做出的假设：较远的东西总是被较近的东西遮挡，也并非一定成立。举个例子，如图 [?] 所示，盘子 (#1) 被咖啡杯 (#5) 遮挡，但咖啡杯在深度上里的更

远。在监督学习方法中，一些工作进行人工顺序标注 [17], [18] 或依赖于人工数据 [16]，通过全监督方法对顺序进行学习。另一个关于前景分割的工作方法，设计了已知端到端的程序来解决片段重叠问题 [26], [27]。然而，这些方法都不能显式恢复场景顺序。

非模态示例分割：模态分割，例如语义分割 [9], [10] 和实例分割 [11]–[13]，目的是为了每个可视像素指定类型或像素标签。现存的模态分割方法无法解决去遮挡问题。不同于模态分割，非模态实例分割旨在检测对象并恢复它们的非模态掩码。李等人 [28] 通过人工添加遮挡物的方式产生虚拟监控，但在遮挡关系复杂且缺少明确的顺序时，分割难度很大。此外其它的工作通过全监督的方式，使用人工标注 [17]–[19] 和人工合成数据 [16] 进行学习。但正如上文提到的，这种方法成本高且在标注不可视掩模时不精确。基于人工数据集的方法面临着真实域和生成域之间的差距问题。相反的，我们的方法可以用自监督方式把模态掩模转化成非模态掩模。这个特殊的能力有助于不进行人工非模态标记就训练出非模态示例分割网络。

非模态补全：非模态补全与模态实例分割有些许不同，在非模态分割中，模态掩模在测试中被给出，任务是完成模态掩模到非模态掩模的转化。先前对非模态掩模补全的研究主要依赖于对不可见边界的启发式假设来实现给定顺序关系的非模态补全。Kimia 等人建议在非模态补全中使用欧拉螺旋 [29]，Lin 等人使用了三次 Bézier 曲线 [30]，Silberman 等人 [31] 利用包括直线和抛物线的曲线原语。因为这些研究都需要顺序作为输入，所以他们不能直接用于解决去障碍问题。除此之外，这些无监督方法主要关注于简单形状的小样本。Kar 等人 [32] 使用关键点注释将 3D 对象模板与 2D 图像对齐，从而生成非模态边界框的真实数据。Ehsani 等人 [15] 利用 3D 合成数据训练端到端非模态补全网络。同无监督方式类似，我们的框架不需要非模态掩码的标注或任何 3D 合成数据。相比之下，我们的方法可以解决高度混乱的自然场景之下非模态补全问题，而其他的无监督方法在这方面有所不足。

III. 我们的场景去遮挡方法

我们提出的框架旨在：1）恢复遮挡顺序 2）补全非模态掩模及被遮挡物的内容。为了解决对顺序及掩模人工标注数据的缺失，我们设计了一种方法来训练我们提出的 PCNet-M 和 PCNet-C，使其以一种自监督的方式部分地完成实例。通过训练好的网络中，我们进一步提出了渐进式推理过程来实现顺序恢复，有序化非模态补全和非模态内容的补全。

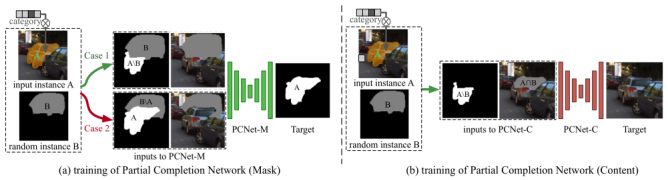

图 3: PCNet-M 和 PCNet-C 的训练过程，以 n 个实例 A 为输入，从整个数据集中随机抽取另一个实例 B，并对其进行随机定位。注意，我们只有 A 和 B 的模态掩码。（A）PCNet-M 是通过切换两种情况来训练的。案例 1（A 被 B 擦除）遵循部分补全机制，鼓励 PCNet-M 部分完成 A。案例 2 防止 PCNet-M 过度补全 A。（B）PCNet-C 使用 $A \cap B$ 擦除 A 并学习填充擦除区域的 RGB 内容。它还接受一个 A\B 作为附加输入。其模式掩码在可用时与其类别 id 相乘。

A. 部分补全网络 *(PCNet)*

在给定图像的前提下，通过现有的实例分割框架获取模态掩模是比较容易做到的，但非模态掩模无法通过该种方式获得。更糟的是，我们无法知道这些模态掩模是否完好无损，这就导致对被遮挡实例的恢复变得极具挑战性。这个问题促使我们探索使用自监督方法部分补全方案。

动机假定一个实例的模态掩模包含像素集 M，用 G 表示真实实例中的非模态掩模。监督学习的方法可以解决 $M \xrightarrow{f_\theta} G$ 的问题，此处的 f_θ 表示整体补全模型。若该实例被多个遮挡物所遮挡，这个整体补全过程可以被拆分成多个部分补全模型的序列：
$M \xrightarrow{f_\theta} M_1 \xrightarrow{f_\theta} M_2 \xrightarrow{f_\theta} \dots \xrightarrow{f_\theta} G$，其中 M_k 是中间状态，P_θ 是部分补全模型。

因为我们仍然没有任何真实数据来训练部分补全模型 P_θ，我们向后退一步，在 M 上随机删减一部分，得到 $M_{-1}, M_{-1} \in M$。然后我们训练模型 P_θ，令 $M_{-1} \xrightarrow{f_\theta} G$，这样的自监督部分补全近似于有监督训

练，是我们 PCNets 的基础。基于这样一个自监督的概念，我们引入了部分补全网络（PCNets）。它包括两个具体网络，针对掩模的 PCNet-M 和针对内容的 PCNet-C。

用于掩模补全的 PCNet-M：对 PCNet-M 的训练过程如图3(a) 所示。我们先准备好训练数据，在有实例级注释给的数据集 D 中，对 PCNet-M 的训练过程如图 3(a) 所示。我们先准备好训练数据，在有实例级注释给的数据集 D 中，给定实例 A 及其掩模 M_A，我们在 D 中随机选取另一个实例 B，将其随机放在一个位置，获得新的掩模 M_B。这里我们把 M_A 与 M_B 看作两个像素集。两个不同的输入样例被放入网络中：

第一个实例对应上述部分补全策略。我们定义 M_B 作为一个擦除器，用 B 擦除 A 的部分内容，获取到 $M_{A\backslash B}$。在该实例中，PCNet-M 被训练来在给定 M_B 的情况下从 $M_{A\backslash B}$ 中恢复原始的掩模 M_A。

第二个实例作为组织网络在没有遮挡的情况下过补全的正则优化项。更具体地说，没有遮挡 A 的 $M_{B/A}$ 被视作擦除器。在这个例子中，我们鼓励 PCNet-M 在有 $M_{B\backslash A}$ 干扰的条件下保留原始的模态掩模 M_A。在没有第二个实例的情况下，PCNet-M 常倾向于增加像素点数，这会导致对于实际没有被遮挡部分的过补全。

在这两个实例中，被擦除的图像块作为补充输入。我们定义损失函数公式如下：

$$L_1 = \frac{1}{N} \sum_{A,B \in D} L(P_\theta^{(m)}(M_{A\backslash B}; M_B, I\backslash M_B), M_A)$$
$$L_2 = \frac{1}{N} \sum_{A,B \in D} L(P_\theta^{(m)}(M_A; M_{B\backslash A}, I\backslash M_{B\backslash A}), M_A)$$

这里的 $P_\theta^{(M)}(*)$ 是我们的 PCNet-M 网络， 代表待优化参数，I 是图像块，L 是二元交叉熵。我们定义最终损失函数 L(m)，这里的 是选择实例 1 的可能性。在两种实例之间的随机选择帮助网络通过两个相邻物体的形状和边界理解他们的顺序关系，进而决定是否去补全该实例。

用于内容补全的 PCNet-C：当我们的目标是补全图像 RGB 内容时，PCNet-C 在思想上与 PCNet-M 一致。正如在图3(b) 中所示，输入实例 A 和 B 和 PCNet-M 一致。在区域 $M_{A\cap B}$ 中的图像像素点被擦除，PCNet-C 的目的即为预测该区域的丢失内容。除此之外，PCNet-C 也利用掩模 A 的剩余部分（即 $M_{A\backslash B}$ 部分）来表示我们补全的是 A 而非其他物体。因此，我们的方法不能简单地用标准图像修复方法替代。

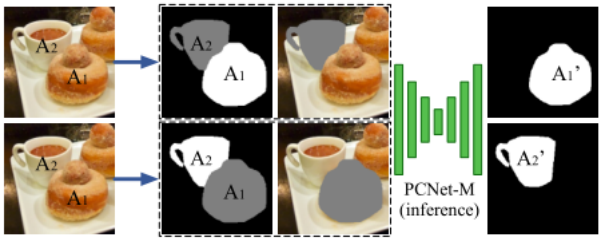

图 4: 用于顺序恢复的双重补全网络。为了恢复香霖实力 A_1 与 A_2 之间的顺序，我们切换目标物体 (白色) 与擦除器 (灰色) 的身份。A_2 的增量大于 A_1 的增量，因此 A_2 被认为是被遮挡物。

PCNet-C 网络试图最小化的损失函数公式如下所示：

$$L^{(c)} = \frac{1}{N} P_\theta^{(c)}(I\backslash; M_{B\backslash A}, I\backslash M_{B\backslash A}), M_A)$$

其中的 $P_\theta^{(c)}$ 是我们的 PCNet-C 网络，I 是图像块，L 代表图像修复中包括感知对抗损失 L_1 在内的的常见损失函数。与 PCNet-M 相似，通过部分补全完成的 PCNet-C 的训练使我们在测试时能够完成实例内容的补全。

B. 用于顺序恢复的双重补全

目标顺序图由所有相邻实例对之间的遮挡关系组成。相邻实例即为两个模态掩模相连的实例，其中一个被另一个遮挡。如图4所示，给定一组相邻实例 A_1 和 A_2，我们首先认为 A_1 的模态掩模 M_{A_1} 是补全目标，M_{A2} 作为获取 A_1 增量（$\Delta_{A_1|A_2}$）的擦除器。对应地，我们也能够获取 A2 在 A1 情况下的增量（($\Delta_{A_2|A_1}$)。在部分补全过程中，被遮挡物会有更大的增量。因此，我们通过 A_1 与 A_2 之间增长区域的比较，推断出了 A_1 与 A_2 间的顺序，公式如下：

$$\Delta_{A_1|A_2} = P_\theta^{(m)}(M_{A_1}; M_{A_2}, I\backslash M_{A_1}),$$
$$\Delta_{A_2|A_1} = P_\theta^{(m)}(M_{A_2}; M_{A_1}, I\backslash M_{A_2}),$$

$$O(A_1, A_2) = \begin{cases} 0 & if|\Delta_{A_1|A_2}| = |\Delta_{A_2|A_1}| = 0 \\ 1 & if|\Delta_{A_1|A_2}| < |\Delta_{A_2|A_1} \\ -1 & \text{otherwise} \end{cases}$$

其中 $O(A_1, A_2) = 1$ 代表 A_1 遮挡 A_2。如果 A1 和 A2 不是相邻的，$O(A_1, A_2) = 0$。注意在实际中 $|\Delta_{A_1|A_2}| = |\Delta_{A_2|A_1}| > 0$ 不成立，因此不需要特别考虑。

对相邻对实现双重补全，我们活得了一张场景遮挡顺序图，该图可以表示为图2之中呈现出的有向图。图中节点代表物体，边代表相邻点对之间遮挡关系。当然，该图不一定是无环的，反例如图7所示。

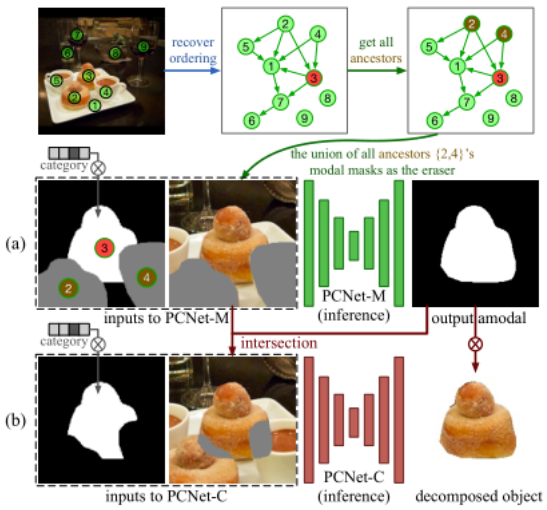

图 5: (a) 有序非模态掩模的补全过程需要目标物体的模态掩模 (#3)，其祖先 (#2,#4) 与被擦除掉的图像作为输入。通过 PCNet-M，得到物体 #3 的非模态掩模。(b) 非模态掩模与其祖先的交集表明物体 #3 的不可视区域。非模态内容的补全 (红色箭头) 被输入 PCNet-C 来为不可视区域填充内容。

C. 非模态与内容补全

基于顺序的非模态补全：在获得顺序图之后，我们可以进行基于顺序做非模态补全。假设我们需要补全实例 A，我们通过 BFS 的方法先在图中找到 A 所有的祖先节点作为该实例的遮挡物。因为图中可能出现环路，所以我们对 BFS 算法做出了相应的调整。有趣的是，我们发现训练好的 PCNet-M 网络实际上可以用到所有的祖先节点作为擦除器。因此，我们不需要遍历其祖先点并逐步应用 PCNet-M 来补全 A。相反，我们可以在先前获得的祖先掩模的基础上进行单步补全。用 $\{Anc_i^A, i = 1, 2, ..., k\}$ 表示 A 的祖先节点，我们按照如下方式进行非模态补全：

$$A_{m_A} = P_\theta^{(m)}(M_A; M_{anc^A}, \backslash M_{anc^A}),$$
$$M_{anc^A} = \bigcup_{i=1}^{k} M_{anc_i^A}$$

这里的 A_{m_A} 是非模态掩模的结果，M_{anc_iA} 是第 i 个祖先节点模态掩模。图5(a) 中展示了一个例子。图6表明了我们使用多个祖先节点而非第一个祖先节点的原因。

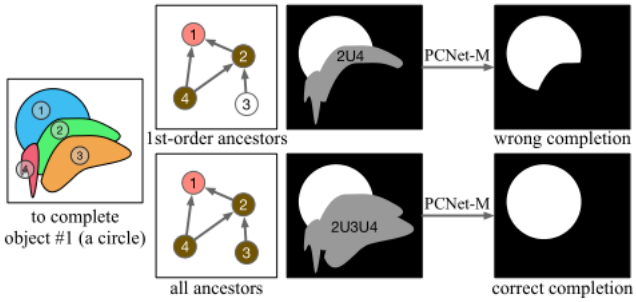

图 6: 这个图显示了为什么我们需要找到所有的祖先而不仅仅是一级祖先，尽管高阶祖先并没有直接遮挡这个实例。高阶祖先 (例如实例 #3) 可能间接遮挡目标实例 (#1)，因此需要考虑。

非模态约束的内容补全：在先前步骤中，我们获得了遮挡顺序图和对每个实例的非模态掩模。接下来，我们要完成被遮挡部分内容的补全。如图5(b) 所示，预测得到的非模态掩模的交集和祖先 $A_{m_A} \cap M_{anc^A}$ 表明了 A 的损失部分，同时被认为是用于 PCNet-C 的擦除器。然后我们使用训练好的 PCNet-C 按照如下方式来填充具体内容：

$$C_A = P_\theta^{(c)}(I \backslash M_E; M_A, M_E) \circ A_{m_A}$$
$$M_E = A_{m_A} \cap M_{anc^A}$$

这里的 C_A 是场景 A 中的分解内容。对于背景内容，我们使用所有前景实例的并集作为擦除器。不同于没有遮挡的图像修复，内容补全只在估计的遮挡区域进行。

IV. 实验

我们在包括顺序恢复，非模态补全，非模态实例分割，场景处理等多种应用中测试了我们的方法，具体实施细节和定性结果可以在补充材料中找到。

数据集. 1) KINS [18] 一个源于 KITTI [20] 的数据集，是一个有模态和非模态标注的大尺度交通数据集。PCNets 在有模态标签的训练集 (7,478 张图像，95,331 个实例) 上进行训练，在测试集 (7,517 张图像，92,492 个实例) 测试去遮挡模型。**2）COCOA** [17] 是 COCO2014 [21] 数据集的一个带有实例对顺序、模态掩模、非模态掩模标注的子集。我们在有模态标签的训

练集 (2,500 张图像，22,163 个实例) 上进行训练，在验证集 (1,323 张图像，12,753 个实例) 测试去遮挡模型。这个数据集实例的类别是无法获得的。因此，我们在这个数据集的训练过程中，置类别标签为 1。

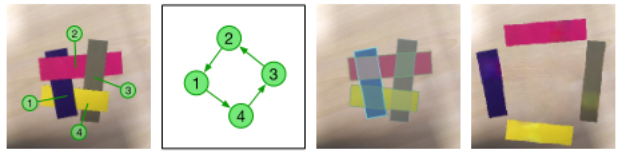

图 7: 我们的框架可以解决循环遮挡问题。由于这样的例子很少，我们剪裁了 4 张纸来形成样例。

表 I: 在 COCOA 验证集合 KINS 测试集上的顺序估计，用相邻实例对之间的顺序计算准确率

method	gt order (train)	COCOA	KINS
Supervised			
$OrderNet^{M}$ [17]	✓	81.7	87.5
$OrderNet^{M+1}$ [17]	✓	81.7	87.5
Unsupervised			
Area	✗	62.4	77.4
Y-axis	✗	58.7	81.9
Convex	✗	76.0	76.3
Ours	✗	**87.1**	**92.5**

A. 结果对比：

顺序恢复：我们在表I中展示了在 COCOA 和 KINS 数据集上顺序恢复的效果。我们复制了在论文 [17] 提出的 OrderNet 获得了有监督的结果。基准包括：根据 [1]，Y 轴排序得到带框实例对（图像下方的实例有限），和凸先验两种。就凸包基准来说，我们在模态掩模上计算凸包来近似非模态补全，有更多增加区域的物体同那个样被认为是被遮挡物。所有的基准都被调整到他们最优的表现状态，在两个基准上，我们的方式都比基准有更优的准确率，达到了基本与监督学习结果相近的程度。在图7中展示了一个四个物体都有循环重叠一个有趣的例子。由于我们的顺序恢复算法回复的是物体对之间的相对顺序而非绝对顺序序列，所以我们能够解决这张图的问题并构建出一个成环有向图。

非模态补全：我们首先介绍基准，对监督方法，我们可以获取非模态标注。AUNet 被训练来端到端地从模态掩模预测非模态掩模。Raw 意味着没有进行补全，Convex 代表着使用凸包方法补全非模态掩模。由于凸包方法通常导致过度补全，即拓展了可视部分的掩模，所以我们通过使用预测顺序来优化凸包，进而优化我们的基准，得到更强的基准 ConvexR，它在自然的凸物体上表现的很好。我们的 (NOG) 代表基于 PCNet-M 得到的无顺序非模态掩模的补全，并且认为其所有的相邻物体均为擦除器而非使用遮挡顺序来搜索祖先，而 (OG) 是有序的非模态补全方法的参考。如表II所示，我们在真实的模态掩模上进行非模态补全。我们的方法优于基准方法，与监督方法基本一致。在 OG 和 NOG 两种方法上的对比，展现了在非模态补全中顺序的重要性。如图9所示，我们的部分结果甚至比人工标注更加自然。除了使用真实模态掩模作为测试时的输入，我们同时用预测出的模态掩模作为输入，验证率我们方法的有效性。特别地，我们训练了一个 UNet 来从图像中预测模态掩模。为了正确地匹配模态和相应的真实非模态掩模，我们使用边界框作为网络中的额外输入。我们在测试集上预测了模态掩模，与真实非模态掩模相比有 52.7 的正确率。我们使用预测得到的模态掩模作为输入非模态补全的输入，如表III所示，即使和监督方法相比，我们的方法同样有很好的表现。

非模态实例分割的标签：非模态实例的分割旨在从那个图像中同时检测实例和预测非模态掩模。通过我们的方法，可以将一个现有的，有模态标注的数据集转换成有伪非模态标注的数据集。该操作通过在模态掩模上训练 PCNet-M 完成，同时如图8所示，在相同的训

表 II: 在 COCOA 验证集合 KINS 测试集上的顺序估计，用相邻实例对之间的顺序计算准确率

method	amodal (train)	COCOA %mIoU	KINS %mIoU
Supervised	✓	82.53	94.81
Raw	✗	65.47	87.03
$Convex^{R}$	✗	74.43	90.75
Ours (NOG)	✗	76.91	93.42
Ours (OG)	✗	**81.35**	**94.76**

表 III: 使用预测得到的掩模 (mAP 52.7 %) 在 KIN 测试集上进行非模态补全)

method	amodal (train)	KINS %mIoU
Supervised	✓	87.29
Raw	✗	82.05
$Convex^{R}$	✗	84.12
Ours (NOG)	✗	85.39
Ours (OG)	✗	**82.26**

表 IV: 基于 KINS 测试集的非模态实例分割，$Convex^R$ 是指用预测的阶数来细化凸包。在这个实验环境中，所有的方法都从原始图像中检测和分割实例。因此，测试中不使用模态掩模。

Ann. source	amodal (train)	amodal (train)	%mAP
GT [18]		✓	29.3
Raw	✓	✗	22.7
$Convex^R$	✓	✗	22.2
Ours (NOG)	✓	✗	25.9
Ours (OG)	✓	✗	**29.3**

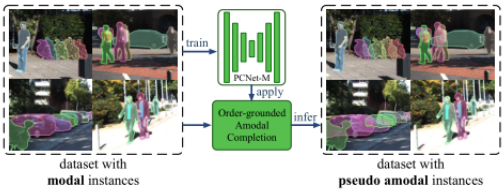

图 8: 通过在模态数据集上训练自监督的 PCNet-M(如这里所示的 KITTI)，并在同一个数据集上应用我们的非模态补全算法，我们能够自由地将模态注释转换为伪非模态注释。注意，这种自监督转换本质上不同于在一个小的标记非模态数据集上训练一个有监督的模型并将其应用于更大的模态数据集，在这种情况下，不同数据集之间的泛化可能存在问题。

练集上应用我们的非模态补全算法来获取非模态掩模。为了衡量伪非模态标签的质量，我们按照论文 [18] 中的设置为非模态实例分割训练了 Mask RCNN 网络 [12]。除了用于训练的非模态标注不同以外，所有基准都遵循相同的训练方法。如表IV所示，通过使用我们推断的非模态边界框和掩模，我们达到了和使用人工非模态标注相同的效果（29.3% 的精度）。除此之外，我们在训练集上推断出的非模态掩模与人工标注的有很高的一致性（95.22%）。该结果表明我们的方法对获取可靠的非模态掩模标注有很高的适用性，这能够减轻人工大型实例数据集标注的负担。

B. 在场景操控中的应用：

我们的场景去遮挡框架允许我们将场景分解为背景和孤立的被补全的物体，同时有一张遮挡顺序图。图片10展现了通过控制顺序合成出的场景，图片11展现出更多的操控实例，表明尽管我们的去遮挡框架的训练与基准相比没有额外信息，但仍然允许我们进行高质量的遮挡操控。

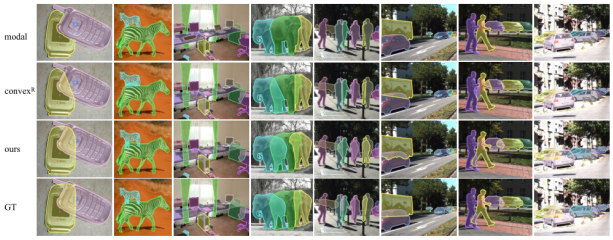

图 9: 非模态补全结果。我们的研究可能比人类标注 (GT) 在一些样例中显得更加自然，尤其是黄色的实例。

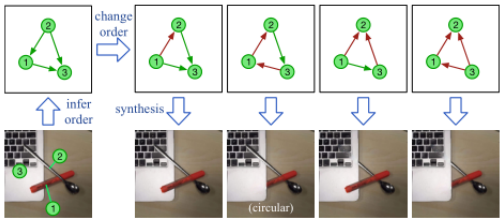

图 10: 通过改变顺序图进行场景合成，相反的顺序用红色箭头显示，也可以合成具有循环序的特殊样例。

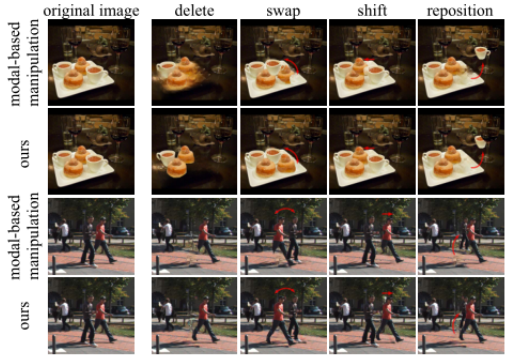

图 11: 此图显示了由我们的方法实现的丰富和高质量的操作，包括删除、交换、移动和重新定位实例上。基于模型的基线方法是基于图像修复的，其中提供了模态掩模，但阶数和模态掩模未知，其在放大视图下效果更好。更多的例子可以在补充材料中找到。

V. 结论

总的来说，我们提出了一个无需顺序或非模态标注的，基于自监督的综合场景的去遮挡框架 PCNet。这个框架以渐进的方式恢复遮挡顺序，然后进行非模态和内容的补全，且在真实数据集与监督学习的方法有相近的正

确率。我们也可以用它将模态标注转化为非模态标注。除此之外，我们的框架能够进行高质量的遮挡场景操控，为图像编辑提供了一个新的维度。

感谢：该工作受到 Sense Time-NTU Collaboration 项目的支持，Sense Time 集团合作研究资助 (CUHK 协议编号 TS1610626& 编号：TS1712093)。

参考文献

[1] Gaetano Kanizsa. Organization in vision: Essays on Gestalt perception. Praeger Publishers, 1979.

[2] Stephen E Palmer. Vision science: Photons to phenomenol- ogy. MIT press, 1999.

[3] Steven Lehar. Gestalt isomorphism and the quantification of spatial perception. Gestalt theory, 21:122–139, 1999.

[4] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. Cascade object detection with deformable part models. In CVPR, pages 2241–2248. IEEE, 2010.

[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[6] Jiaqi Wang, Kai Chen, Shuo Yang, Chen Change Loy, and Dahua Lin. Region proposal by guided anchoring. In CVPR, 2019.

[7] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuan- jun Xiong, Chen Change Loy, and Dahua Lin. Optimizing video object detection via a scale-time lattice. In CVPR, June 2018.

[8] Kai Chen, Jiaqi Wang, Shuo Yang, Xingcheng Zhang, Yuan- jun Xiong, Chen Change Loy, and Dahua Lin. Optimizing video object detection via a scale-time lattice. In CVPR, June 2018.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolu- tion, and fully connected crfs. PAMI, 40(4):834–848, 2017.

[10] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In CVPR, pages 2881–2890, 2017.

[11] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In ECCV, pages 534–549. Springer, 2016.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Gir- shick. Mask r-cnn. In ICCV, 2017.

[13] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaox- iao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance seg- mentation. In CVPR, pages 4974–4983, 2019.

[14] Jiaqi Wang, Kai Chen, RuiXu, ZiweiLiu, ChenChangeLoy, and Dahua Lin. Carafe: Content-aware reassembly of fea- tures. In ICCV, October 2019.

[15] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In CVPR, pages 6144–6153, 2018.

[16] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. Sail-vos: Semantic amodal in- stance level video object segmentation-a synthetic dataset and baselines. In CVPR, pages 3105–3115, 2019.

[17] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In CVPR, pages 1464–1472, 2017.

[18] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In CVPR, pages 3014–3023, 2019.

[19] Patrick Follmann, Rebecca Kö Nig, Philipp Hä Rtinger, Michael Klostermann, and Tobias Bö Ttger. Learning to see the invisible: End-to-end trainable amodal instance segmen- tation. In WACV, pages 1328–1336. IEEE, 2019.

[20] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. The Inter- national Journal of Robotics Research, 32(11):1231–1237, 2013.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, and Piotr Dollár. Microsoft coco: Common objects in context. In ECCV, 2014.

[22] Jiajun Wu, Joshua B Tenenbaum, and Pushmeet Kohli. Neu- ral scene de-rendering. In CVPR, 2017.

[23] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene parsing with object instances and occlusion ordering. In CVPR, pages 3748–3755, 2014.

[24] Derek Hoiem, Andrew N Stein, Alexei A Efros, and Mar- tial Hebert. Recovering occlusion boundaries from a single image. In ICCV, 2007.

[25] Pulak Purkait, Christopher Zach, and Ian Reid. Seeing be- hind things: Extending semantic segmentation to occluded regions. arXiv preprint arXiv:1906.02885, 2019.

[26] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In CVPR, pages 6172–6181, 2019.

[27] Justin Lazarow, Kwonjoon Lee, and Zhuowen Tu. Learning instance occlusion for panoptic segmentation. arXiv preprint arXiv:1906.05896, 2019.

[28] Ke Li and Jitendra Malik. Amodal instance segmentation. In ECCV, pages 677–693. Springer, 2016.

[29] Benjamin B Kimia, Ilana Frankel, and Ana-Maria Popescu. Euler spiral for shape completion. IJCV, 54(1-3):159–182, 2003.

[30] Hongwei Lin, Zihao Wang, Panpan Feng, Xingjiang Lu, and Jinhui Yu. A computational model of topological and geo- metric recovery for visual curve completion. Computational Visual Media, 2(4):329–342, 2016.

[31] Nathan Silberman, Lior Shapira, Ran Gal, and Pushmeet Kohli. A contour completion model for augmenting surface reconstructions. In ECCV, 2014.

[32] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jiten- dra Malik. Amodal completion and size constancy in natural scenes. In ICCV, pages 127–135, 2015.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U- net: Convolutional networks for biomedical image segmen- tation. In International Conference on Medical image com- puting and computer-assisted intervention, pages 234–241. Springer, 2015.

[34] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for ir- regular holes using partial convolutions. In ECCV, 2018.

VI. A. 附加材料

在我们的实验中，PCNet-M 的主干网络是加宽系数为 2 的 UNet 网络 [33]，而 PCNet-C 是一个增加了部分卷积层 [34] 的 UNet 网络。但注意，PCNet 对于主干网络并无限制。对这两种 PCNet，以物体为中心的图像或掩模块被一个自适应方块剪裁并重塑为 256*256 的输入。　　对于 COCOA 数据集，PCNet-M 使用 SGD 进行 56K 次训练，初始学习率为 0.001，且在 32K 次和 48K 的训练中衰减到 0.1。对 KINS 数据集，我们在 32K 次之前停止了训练进程。批处理量为 256，分布在在 8 块 GTX1080Ti 上进行计算。平衡训练 PCNet-M 网络中两种实例的超参数　置为 0.8。在最近的实验中，我们不采用 RGB 色域的图片作为 PCNet-M 的输入，因为我们从经验上发现：通过串联引入 RGB 并没有很好的优化。这可能是因为对这两个数据集，模态掩模的信息量对训练已经足够；但我们相信在更复杂的场景中，RGB 的引入将发挥更好的作用。

对于 PCNet-C，我们修改了 UNet，将图像和模态掩码的级联作为输入。除了 [34] 中的损失 [34] 外，我们还增加了额外的对抗性损失，以进行优化。该判别器由 5 个卷积层叠加而成，同时应用了谱归一化以及 leaky Relu(斜率为 0.2)。PCNet-C 经过微调，可进行 450K 迭代，并保持从预训练网络 [34] 中获得 $10^{-}4$ 的恒定学习速率 [34]。我们应用预训练的权重，以适应接受额外的模态掩模。

VII. B. 讨论

B.1. 不同遮挡率的分析

图12展示了在不同方法不同遮挡率下非模态补全的效果，总的来说，大的遮挡率会导致较低的表现效果。在高遮挡率的情况下，我们所有的方法（OG）都有超过基线很高的表现。

B.2. 该方法是否支持互遮挡？

作为一个缺点，因为我们的方法侧重于对象级别的去遮挡，我们的方法不支持如图13所示的互相遮挡的情况。对于相互遮挡情况，顺序图不能被定义，因此需要精细的边界层次的去遮挡。尽管如此，如主论文图 7 所示，如果有两个以上的物体有循环遮挡情况，我们的方法可以有很好的效果。

B.3. 实例 2 是否会误导 PCNet-M?

如图14所示，人们会关心当在实例 (a-2) 应用不完整策略时，A 和 $B\backslash A$ 的边界可能包括 A 被实际物体阻挡的绿色的区域。因此，如果黄色阴影区域被认为为不填充，可能对 PCNet-M 造成误导。

这里我们对为什么不会造成误导进行解释：首先，PCNet-M 学习在有遮挡的情况下补全或不补全目标物体。如图14所示，由于 PCNet-M 被训练着　　在（A-1）中补全 $A\backslash B$，而不是在（A-2）补全 A，所以有必要发现在 (a-1) 中 A 在 B 之下且 (a-2) 中 A 在 B 之上的证据。该证据包括两个物体的形状、共同边界的形状、连接等。在测试时，如（b）中以真实 C 为条件时，PCNet-M 很容易从这些线索中判断 C 在 A 之上。因此当 A 将在 C 的条件下被补全时，PCNet-M 实际上倾向于实例 1。

那么这对完成策略有什么影响呢？（c）中的情况与 (a-2) 具有非常相似的遮挡模式，特别是在公共边界的右上部分，显示了 a 在 c 之上的强烈线索，在这种情况下，PCNet-M 将无法按预期完成 a。然而，案例（c）是不正常的，它不可能存在于现实世界中。未完成策略真正生效的情况是案例（d），在这种情况下，当强烈的线索表明 A 在 D 之上时，PCNet-M 被训练着不要越过 A 和 D 的边界来入侵 D。

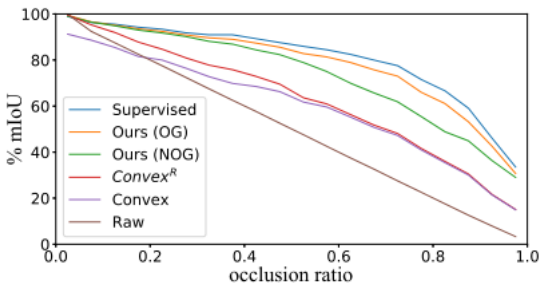

图 12: 在 KINS 测试装置上评价了不同入路在不同遮挡比下的性能。

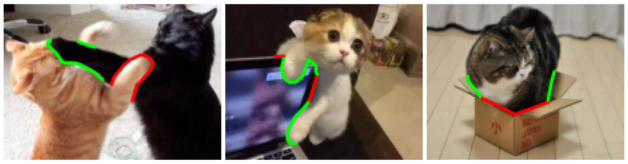

图 13: 相互遮挡样例，绿色边界表示一个对象遮挡另一个对象，红色边界则相反。

VIII. C. 可视化

如图15所示，我们的方法使我们能够自由地调整场景的空间参数来重新排布新场景。其质量可以通过好的图像复原方法提升，因为 PCNet-C 和图像复原共享着一个相似的网络结构与训练策略。

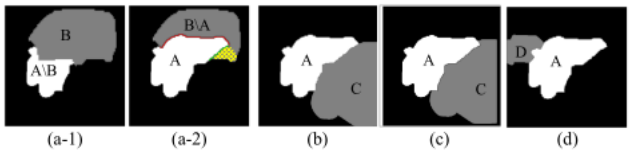

图 14: （a-1）和（a-2）分别代表培训中的案例 1 和案例 2；（b）-（d）代表测试中可能的案例。在这些测试用例中，只补全 (b) 中的 A。

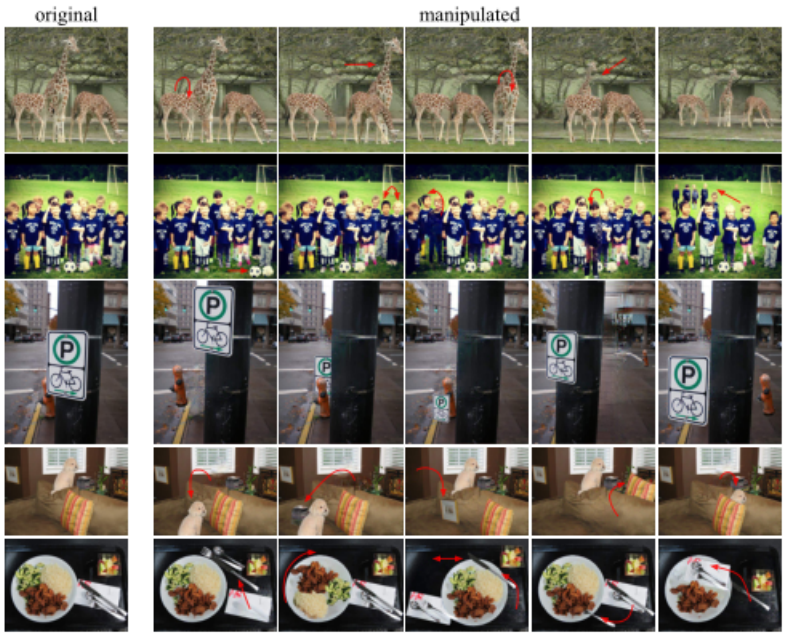

图 15: 基于我们的去遮挡框架的场景操作结果。不明显的变化用红色箭头标记。视频演示可以在项目页面中找到